Cumulation of poverty measures to meet new policy needs

A.Lemmi, V.Verma, G.Betti, L.Neri, F.Gagliardi, G.Tarditi and C.Ferretti
Table of contents

1. Context and scope
2. Gain from cumulation over waves
3. Variance estimation for complex statistics
4. Software development
Context and scope - 1

- Reliable indicators of poverty and social exclusion are an essential monitoring tool. In the EU-wide context, these indicators are most useful when they are comparable across countries and over time.

- Furthermore, policy research and application require statistics **disaggregated to increasingly lower levels and smaller subpopulations**. Direct, one-time estimates from surveys designed primarily to meet national needs tend to be insufficiently precise for meeting these new policy needs.
Context and scope - 2

• This is particularly true in the domain of poverty and social exclusion, the monitoring of which requires complex distributional statistics – statistics necessarily based on intensive and relatively small-scale surveys of households and persons.

• This paper addresses some statistical aspects relating to improving the sampling precision of such indicators for sub-national regions in EU countries, in particular through the *cumulation of data over rounds of regularly repeated national surveys*. 
EU-SILC - 1

- The reference data for this purpose are EU Statistics on Income and Living Conditions, the major source of comparative statistics on income and living conditions in Europe. A standard integrated design has been adopted by nearly all EU countries.

- It involves a rotational panel, with a new sample of households and persons introduced each year to replace one-fourth of the existing sample. Persons enumerated in each new sample are followed-up in the survey for four years. The design yields each year a cross-sectional sample, as well as longitudinal samples of 2, 3 and 4 year duration.
EU-SILC - 2

SAMPLE

<table>
<thead>
<tr>
<th>TIME</th>
<th>T-3</th>
<th>T-2</th>
<th>T-1</th>
<th>T</th>
<th>T+1</th>
<th>T+2</th>
<th>T+3...</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Years in survey
Different modes of using data

Survey data such as from EU-SILC can be used in different forms to construct regional indicators.

(1) Direct estimation from survey data – in the same way as at the national level – provided that the regional sample sizes are large enough.

(2) Constructing alternative indicators - but with substantively similar meaning - which utilise the available survey data more intensively.

(3) **Cumulation of data over survey waves to increase precision of the direct estimates.**

(4) Using survey data in conjunction with data from other - especially administrative - sources, to produce improved estimates using small area estimation (SAE) techniques.
Gain from cumulation over waves -1

Consider that a person’s poverty status (poor or non-poor) is determined from the income distribution for each wave separately, and the proportion of poor at that wave is computed. These proportions are then averaged over a number of consecutive waves.

The issue is to quantify the gain in sampling precision from such pooling, given that data from different waves of a rotational panel are highly correlated.
Gain from cumulation over waves - 2

Under a two waves model and full overlap, the variance of the average can be estimated as:

\[ v_A = \frac{V}{2} (1 + b) \]

Where:
- \( V \) is the cross-sectional variance of the poverty rate
- \( b = \left( \frac{c_1}{v} \right) = \left( \frac{a - p^2}{p - p^2} \right) \)

Alternatively, if the overlap is only partial:

\[ v_A = \frac{1}{2} \frac{V_1 + V_2}{2} (1 + b \left( \frac{n}{n_h} \right)) \]

Where:
- \( a \) is the persistent poverty rate over the two years
- \( n \) is the overlap between the cross-sectional samples
- \( n_h \) is the harmonic mean of the cross-sectional sample size \( n_1 \) and \( n_2 \)
Gain from cumulation over waves - 3

In the case of multiple waves (full overlap) the correlation will be inversely proportional to distance between years, hence we can assume:

\[ \frac{c_k}{v} = \left( \frac{c_1}{v} \right)^k \]

From which we can derive the gain in precision averaging over K waves.

\[ \frac{\nu_k}{\nu} = \frac{1}{K} \left( 1 + 2 \sum_{k=1}^{K-1} \frac{(K-k)}{K} \left( \frac{c_1}{v} \right)^k \right) \]
Assumption in our work

Given that in EU-SILC data cross-sectional datasets cannot be linked, correlation between the two waves cannot be computed directly.

We computed it through the longitudinal dataset. So:

\[ b = \left( \frac{a - p^2}{p - p^2} \right) \]

where \( a \) is persistent poverty rate of the panel (e.g. 2007-2008 for Italy) and \( p \) is the average of the estimates done in the longitudinal dataset for each of the two years (e.g. 2007-2008 for Italy).

\[ n_h = \frac{2 \cdot n_1 \cdot n_2}{n_1 + n_2} \]

where \( n_1 \) and \( n_2 \) are the sample size of the cross-sectional dataset.

\[ n = 0.75 \cdot \min(n_1, n_2) \cdot \frac{m}{\min(m_1, m_2)} \]

where \( m \) is the sample size of the panel (e.g. 2007-2008 for Italy) and \( m_1 \) and \( m_2 \) are the sample size of the longitudinal dataset in the two years.
Gain from cumulation over two waves

<table>
<thead>
<tr>
<th>S.e. of average HCR</th>
<th>Italy 07/08</th>
<th>Poland 05/06</th>
</tr>
</thead>
<tbody>
<tr>
<td>(assuming independent samples)</td>
<td>(1) = \left(\frac{v_1 + v_2}{2}\right)^{0.5}</td>
<td>0.36</td>
</tr>
<tr>
<td>Factor by which s.e. is increased</td>
<td>(2) = \left(1 + b\left(\frac{n}{n_h}\right)\right)^{0.5}</td>
<td>1.20</td>
</tr>
<tr>
<td>(due to positive correlation between waves)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S.e. Of average HCR</td>
<td>(3) = (1)(2)</td>
<td>0.43</td>
</tr>
<tr>
<td>(over two years, given correlated samples)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average s.e.</td>
<td>(4) = \left(\frac{(v_1)^{0.5} + (v_2)^{0.5}}{2}\right)</td>
<td>0.50</td>
</tr>
<tr>
<td>(over a single year)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average gain in precision</td>
<td>(5) = 1 - (\frac{3}{4})^2 \times 100%</td>
<td>26%</td>
</tr>
<tr>
<td>(variance reduction, or increase, in effective sample size over a single year sample)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Variance estimation for complex statistics - 1

The Jackknife Repeated Replication (JRR) provides a versatile technique for variance estimation for complex statistics such as cumulative and longitudinal measures. JRR is one of the variance estimation methods based on comparisons among replications generated through repeated re-sampling of the parent sample. Once the set of replications has been defined for any complex design, the same variance estimation algorithm applies to a statistic of any complexity:

\[ \text{var}(Q) = \sum_k \left[ (1 - f_k) \frac{a_k - 1}{a_k} \sum_j (Q_{kj} - Q_k)^2 \right] \]
Variance estimation for complex statistics - 2

In the standard application of JRR, a replication is formed by:
1. eliminating one PSU from the sample
2. appropriately increasing the weight given to remaining units in stratum.

The number of replications generated is the same as the number of PSUs in the sample.
Variance estimation for complex statistics - 3

We have applied this method for estimating variances in the EU-SILC rotational panel design for:

1. Sub-populations (including geographical domains)
2. Longitudinal measures such as persistent poverty rates
3. Measures of net change and averages over cross-sections
Variance estimation for complex statistics - 4

The total sample of interest is formed by the union of all cross-sectional samples being compared or aggregated.

Using as basis the common structure of this total sample, a set of JRR replications is defined in the usual way.

Each replication is formed such that when a unit is to be excluded in its construction, it is excluded simultaneously from every wave where the unit appears.
For each replication, the required measure is constructed for each of the cross sectional samples involved, and these measures are used to obtain the required averaged measure for the replication.

The variance of the statistic of interest is estimated from the replication estimates in the usual way.
SAS and R programs have been implemented for the estimation. The logic of the software is the following:

1. Estimate from the *longitudinal dataset*: a and m (for the two years panel), p1, p2, m1 and m2.

2. Estimate from the two *cross sectional datasets* (once the appropriate structure, STRATA and PSUs, are constructed): est1, s.e.1, n1, est2, s.e.2, n2.
Concluding Remarks - 1

Appropriate coding of the sample structure, in the survey micro-data and accompanying documentation, is an essential requirement in order to compute sampling errors taking into account the actual sample design.

Lack of information on the sample structure in survey data files is a long-standing and persistent problem in survey work, and unfortunately affects EU-SILC as well. Indeed, the major problem in computing sampling errors for EU-SILC is the lack of sufficient information for this purpose in the micro-data available to researchers.
In many cases the limited information prevents direct and complete computation of variances. We have developed *approximate procedures* in order to reduce these limitations, and used them to produce useful estimates of sampling errors.
Decomposition of variances and design effects identifies more ‘portable’ components, which may be more easily imputed (carried over) from a situation where they can be computed with the given information, to another situation where such direct computations are not possible. On this basis valid estimates of variances can be produced for a wider range of statistics, thus at least partly overcoming the problem due to lack of information on sample structure.
Software Development

Statistic Software: SAS and R

- Two different generations
- How to evaluate their performance
- Difficulties with discussing code properties
Software Development

Why add R:

- Functionality
- Flexibility
- Sharing
- Costs
- License
Software Development

The SAS Code:

- General structure
- DATA and PROC steps
- Sub-structure operations
- Comparison problems
- Computation time issues
Software Development
References